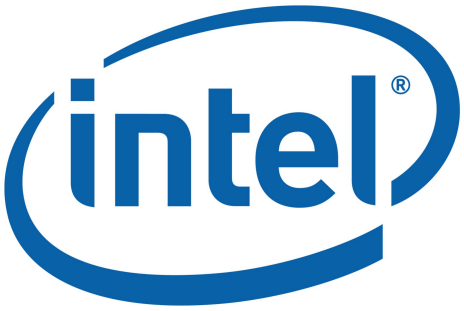


Intel® Ethernet Switch FM4000

24-Port 10G Ethernet L2/L3/L4 Switch/Router

Specification Update

Networking Division (ND)

January 2014

Revision 1.1

LEGAL

Revision History

Revision	Date	Comments
1.1	January 24, 2014	**Documentation Updates added or updated:** • 1. REFCLK Input Characteristics Added to Datasheet (Added) • 2. Datasheet Converted to New Document Template (Added) • 3. Specification Update Converted to New Document Template (Added) **Other Updates:** • Product Code, Device Identification, Component Markings and MM Numbers added to Section 1.0 on the *Intel® Ethernet Switch FM4000 Specification Update*.
1.0	May 21, 2013	Initial release. (Intel confidential)

NOTE: ***This page intentionally left blank.***

1.0 Introduction

This document applies to the FM4000.

1.1 Product Code and Device Identification

Product Code:

The following tables and drawings describe the various identifying markings on each device package:

Table 1-1 Markings

Device	Stepping	Top Marking	Q-Specification	Description
Intel® Ethernet Switch FM4105	A3	EZFM4105F897C	SLKAM	2 10 GbE ports and 16 1 GbE ports, fully-integrated, single-chip wire-speed, layer-2/3/4, 10 GbE/ 2.5 GbE/1 GbE Ethernet switch.
	A3	EZFM4105F897E	SLKAN	
Intel® Ethernet Switch FM4112	A3	EZFM4112F897C	SLJMZ	8 10 GbE ports and 16 1 GbE ports, fully-integrated, single-chip wire-speed, layer-2/3/4, 10 GbE/ 2.5 GbE/1 GbE Ethernet switch.
	A3	EZFM4112	SLJMZ	
	A3	EZFM4112	SLJPB	
	A3	EZFM4112	SLJPC	
	A3	EZFM4112F897E	SLJPB	
Intel® Ethernet Switch FM4212	A3	EZFM4212F1433E	SLJMX	12 10 GbE port, fully-integrated, single-chip wire-speed, layer-2/3/4, 10 GbE/2.5 GbE/1 GbE Ethernet switch.
	A3	EZFM4212F1433C	SLJMW	
Intel® Ethernet Switch FM4224	A3	EZFM4224F1433E	SLKAD	24-port, fully-integrated, single-chip wire-speed, layer-2/3/4, 10 GbE/ 2.5 GbE/1 GbE Ethernet switch.
Intel® Ethernet Switch FM4410	A3	FBFM4410F529E	SLJM7	24-port, fully-integrated, single-chip wire-speed, layer-2/3/4, 10 GbE/ 2.5 GbE/1 GbE Ethernet switch.
	A3	FBFM4410F529C	SLJM6	

Table 1-2 MM Numbers

Product	Tray MM#	
Intel® Ethernet Switch FM4105	930439 930440	
Intel® Ethernet Switch FM4112	920720 920721 920723 921071 921072	
Intel® Ethernet Switch FM4212	921070 921073	
Intel® Ethernet Switch FM4224	930428	
Intel® Ethernet Switch FM4410	920987 920988	

1.2 Marking Diagrams

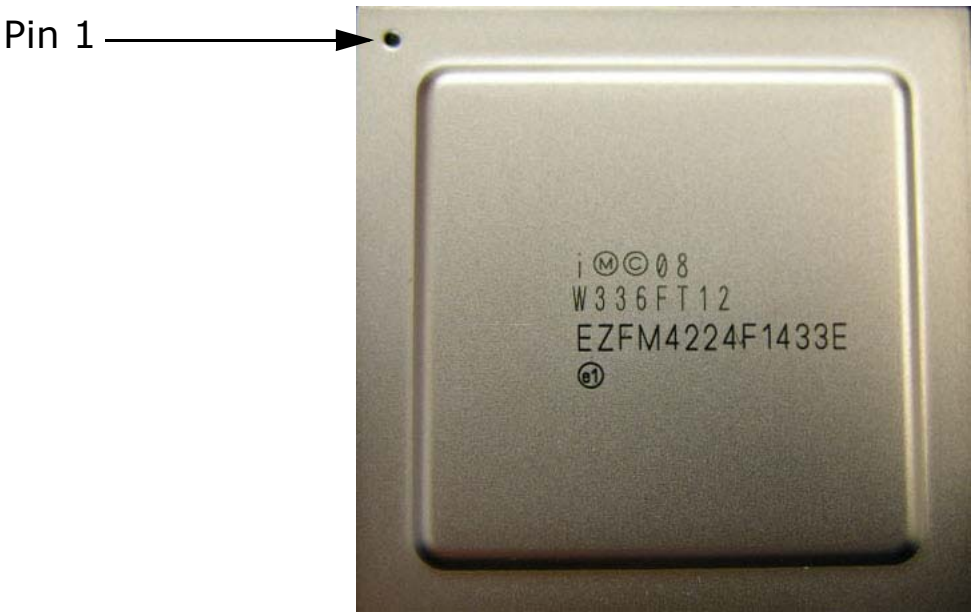
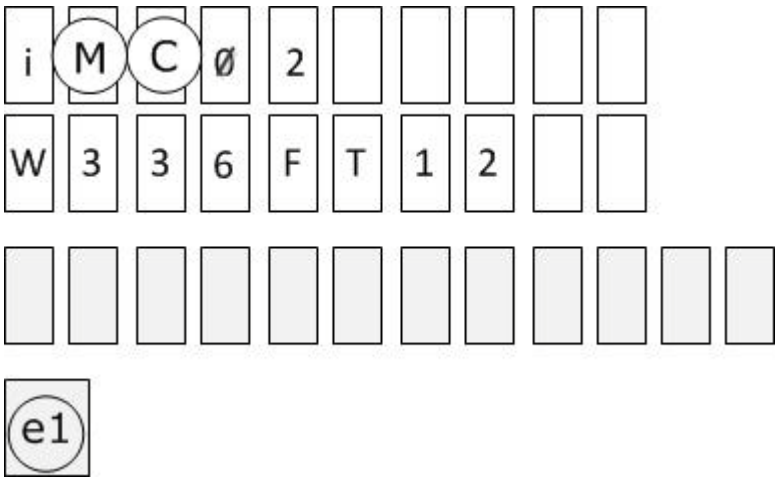


Figure 1-1. Example With Identifying Marks



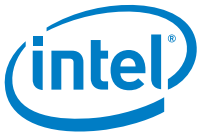
- LINE1: i, maskwork, and copyright
- LINE2: FPO number
- LINE3: Product name followed by test designator (C, E, or I). Left blank by assembly sub-contractor.
- LINE4: Pb free symbol

1.3 Nomenclature Used in This Document

This document uses specific terms, codes, and abbreviations to describe changes, errata, sightings and/or clarifications that apply to silicon/steppings. See Table 1-1 for a description.

Table 1-1 Nomenclature

Name	Description
Specification Clarifications	Greater detail or further highlights concerning a specification's impact to a complex design situation. These clarifications will be incorporated in the next release of the specifications.
Specification Changes	Modifications to the current published specifications. These changes will be incorporated in the next release of the specifications.
Errata	Design defects or errors. Errata may cause device behavior to deviate from published specifications. Hardware and software designed to be used with any given stepping must assume that all errata documented for that stepping are present on all devices.
Software Clarifications	Applies to Intel drivers, EEPROM loads.
Documentation Changes	Typos, errors, or omissions from the current published specifications. These changes will be incorporated in the next release of the specifications.
A1, B1, etc.	Stepping to which the status applies.
Doc	Document change or update that will be implemented.
Fix	This erratum is intended to be fixed in a future stepping of the component.
Fixed	This erratum has been previously fixed.
NoFix	There are no plans to fix this erratum.
Eval	Plans to fix this erratum are under evaluation.
Red Change Bar/or Bold	This Item is either new or modified from the previous version of the document.



2.0 Hardware Clarifications, Changes, Updates and Errata

See Section 1.3 for an explanation of terms, codes, and abbreviations.

Table 2-1 Summary of Specification Clarifications

Specification Clarification	Status
None	N/A

Table 2-2 Summary of Specification Changes

Specification Change	Status
None	

Table 2-3 Summary of Documentation Updates

Specification Change	Status
1. REFCLK Input Characteristics Added to Datasheet	N/A
2. Datasheet Converted to New Document Template	N/A
3. Specification Update Converted to New Document Template	N/A

Table 2-4 Summary of Errata; Errata Include Steppings

Erratum	Status
1. “check end” Check Not Performed	A0-A3=Yes; NoFix
2. Default Rate Limiter Settings Limit Minimum Frame Handler Frequency	A0-A3=Yes; NoFix
3. Frames May Be Dropped with a High Volume of Mixed Traffic Types	A0-A3=Yes; NoFix
4. Hardware Purge of MAC Table Does Not Work as Expected	A0-A3=Yes; NoFix

Table 2-4 Summary of Errata; Errata Include Steppings (Continued)

Erratum	Status
5. No Way to Disable CM_SHARED_PAUSE_WM per Port	A0-A3=Yes; NoFix
6. Canonical GloRT is Applied; GloRT for the Member Port is Expected	A0-A3=Yes; NoFix
7. Double Tagging Identical VLAN Ethertypes Breaks L3/L4 Parsing	A0-A3=Yes; NoFix
8. TCN Precedence is Incorrect for Learning or Aging Events	A0-A3=Yes; NoFix
9. Strict/Round-Robin Scheduler Bug	A0-A3=Yes; NoFix
10. RX Rate Limiter Drop Threshold Quanta	A0-A3=Yes; NoFix
11. tcnCountErrorEvents Becomes Invalid, Blocking New Error Events	A0-A3=Yes; NoFix
12. CPU May Encounter Very Long Bus Cycles When Sending Frames to the Switch	A0-A3=Yes; NoFix
13. Reserved Frames Similar to PAUSE Are Not Dropped Correctly	A0-A3=Yes; NoFix
14. IP Multicast Packets Are Counted as Flood Packets in Group 6	A0-A3=Yes; NoFix
15. SGMII Auto-Negotiation Reports Failure Even When Successful	A0-A3=Yes; NoFix
16. Parity Error Interrupts Are Not Reported in LSM_INT_DETECT	A0-A3=Yes; NoFix
17. CRC Check Not Done on Truncated Frames in Cut-Through Mode	A0-A3=Yes; NoFix
18. EBI Timing Specification Needs to Be Changed	A0-A3=Yes; NoFix
19. Parse Errors Cannot Be Disabled	A0-A3=Yes; NoFix
20. TCN FIFO Has MA_TABLE Parity Error Bit Set When Soft Learning is On	A0-A3=Yes; NoFix
21. Multi-Chip TX Mirroring Requires Dedicated Inter-Switch Link on Mirroring Switch	A0-A3=Yes; NoFix
22. The Datasheet Description of Register CM_GLOBAL_WM is Insufficient	A0-A3=Yes; NoFix
23. TX Polarity Reversal is Per-Port Not Per-Lane	A0-A3=Yes; NoFix
24. Idle Disparity Problem with Lane Reversal in 1 GbE mode	A0-A3=Yes; NoFix
25. MTU Checks Applied to All Unicast IP Frames	A0-A3=Yes; NoFix
26. IP Option Traps Are Applied to All IP Frames	A0-A3=Yes; NoFix
27. Multiple Triggers with Action 'Drop' Cannot Be Combined	A0-A3=Yes; NoFix
28. In 10/100/1000 Mode, Non Word Aligned Frames May Stretch IFG	A0-A3=Yes; NoFix
29. Certain FIBM Request Message Causes Runt Response Message	A0-A3=Yes; NoFix
30. Untagged VLAN Tag Frames Do Not Get VID/VPRI Values in ISL Tag	A0-A3=Yes; NoFix
31. CPU Port Does Not Pad Certain Frame Sizes During Frame Transmission	A0-A3=Yes; NoFix

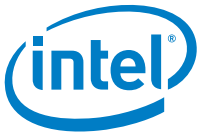


Table 2-4 Summary of Errata; Errata Include Steppings (Continued)

Erratum	Status
32. Management Access to FID Tables Not Reliable Under Heavy Traffic Load	A0-A3=Yes; NoFix
33. An FFU TCAM Read Can Corrupt a Subsequent FFU TCAM Write	A0-A3=Yes; NoFix
34. Erroneous Port Linkup in 1 GbE Mode	A0-A3=Yes; NoFix
35. High Volume of Frame Discards in the Scheduler May Cause the Device to Become Unresponsive	A0-A3=Yes; NoFix

Table 2-5 Summary of Fixed Errata; Errata Include Steppings

Fixed Erratum	Status
1. All /R/ Columns Are Deleted	A0-A2=Yes; NoFix. A3=No; Fixed
2. LCI RX_RDY De-Asserted One Cycle Late	A0-A1=Yes; NoFix. A2=No; Fixed
3. Statistics Gathering May Lead to Deadlock Under Heavy Load	A0-A1=Yes; NoFix. A2=No; Fixed
4. Source Address Lookup May Lead to Deadlock Under Heavy Load	A0-A1=Yes; NoFix. A2=No; Fixed
5. FM4000 May Be Irreparably Damaged with Improper Bring-Up Sequence	A0-A1=Yes; NoFix. A2=No; Fixed
6. Pause Reception May Operate Unreliably with Routing Enabled	A0-A1=Yes; NoFix. A2=No; Fixed
7. Pause Transmission May Operate Unreliably	A0-A1=Yes; NoFix. A2=No; Fixed
8. MAC Address Table Change Notifications May Cease at High Temperature	A0-A2=Yes; NoFix. A3=No; Fixed
9. Source MAC Address Migration May Not Be Reported Reliably	A0-A1=Yes; NoFix. A2=No; Fixed
10. Background MAC Address Table Management Access May Freeze Aging	A0-A1=Yes; NoFix. A2=No; Fixed
11. High Level of Transitions in Frame Handler May Cause Unpredictable Behavior	A0-A1=Yes; NoFix. A2=No; Fixed
12. Tx/Rx Queue Bottleneck May Lead to Deadlock	A0-A2=Yes; NoFix. A3=No; Fixed
13. SPI MOSI/MISO Pins Reversed (897-Ball Package Only)	A0-A2=Yes; NoFix. A3=No; Fixed
14. Triggered Forced Learning Not Operational	A0-A1=Yes; NoFix. A2=No; Fixed
15. Group 4 and Group 5 Counters Are Unreliable	A0-A1=Yes; NoFix. A2=No; Fixed
16. PARITY_EVEN Needs Pull-Up or Pull-Down	A0-A1=Yes; NoFix. A2=No; Fixed
17. VLAN Counters Do Not Clear on Write	A0-A2=Yes; NoFix. A3=No; Fixed
18. Power-Up Supply Sequence Must Be Observed	A0-A2=Yes; NoFix. A3=No; Fixed

2.1 Specification Clarifications

None.

2.2 Specification Changes

None.

2.3 Documentation Updates

1 REFCLK Input Characteristics Added to Datasheet

Change:

A table showing REFCLK input characteristics was added to Section 15.0, "Electrical Specification" in the *Intel® Ethernet Switch FM4000 Datasheet*.

2 Datasheet Converted to New Document Template

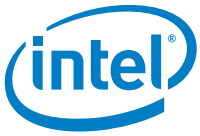
Change:

The *Intel® Ethernet Switch FM4000 Datasheet* was updated to a new document template to comply with Intel corporate standards.

3 Specification Update Converted to New Document Template

Change:

The *Intel® Ethernet Switch FM4000 Specification Update* (this document) was updated to a new document template to comply with Intel corporate standards.



2.4 Errata

1 "check end" Check Not Performed

Problem:

The IEEE specification calls for a check known as "check end", which is designed to detect disparity errors that have been propagated beyond the data payload and into the idle characters following the frame. Two optional modes are implemented for the FM4000:

- Mode 1: The check is not performed
- Mode 2: The check is performed

The first operating mode (Mode 1) is consistent with earlier implementations. However, it is not fully compliant with the UNH test suite. This is not a severe problem; UNH tests for this condition, but dismisses the failure mode as irrelevant. The probability of a "check end" error being generated and then, in the same packet, the packet error being missed by the CRC check is infinitesimally small, making it unlikely to ever occur during the lifetime of the device.

The second operating mode (Mode 2) does not operate correctly. Furthermore, it is the default mode on the chip. Thus, the device must be actively placed on Mode 1.

Implication:

Mode 2 should not be used.

When Mode 1 is enabled, there is a remote chance that a "check end" error occurs and, in the same packet, the packet error is not caught by the CRC check.

Workaround:

The FM4000 should be placed into Mode 1 for normal operation.

The API has been modified to override the default value (Mode 2) with the appropriate value (Mode 1). If customers are not using the API software, they should implement the same capability in their API.

Status: A0-A3=Yes; NoFix

2 Default Rate Limiter Settings Limit Minimum Frame Handler Frequency

Problem:

The default values of RX_RATE_LIM_CFG.RateUnit and TX_RATE_LIM_CFG.RateUnit allow the Frame Handler (FH) clock frequency to be reduced to as low as 280 MHz while maintaining 10 Gb/s per-port data-rate. To reduce the FH frequency further, the RateUnits in both of these registers need to be set to 0x7f. This allows the FH frequency to be reduced down to the minimum-supported 246 MHz without inadvertently rate-limiting the ports (i.e., at a 246 MHz FH frequency, any given port can operate up to 10 Gb/s, sustained).

Note: Reducing the FH frequency below 246 MHz is currently not supported.

Implication:

Reducing the FH frequency helps avoid the unpredictable behavior described in erratum 3 below. However, the FH frequency cannot be reduced until after the rate limiter settings are changed appropriately, as described in this erratum.

Workaround:

The RX_RATE_LIM_CFG.RateUnit and TX_RATE_LIM_CFG.RateUnit registers can be manually set to 0x7f.

The default settings are appropriate under normal operating conditions.

Status: A0-A3=Yes; NoFix

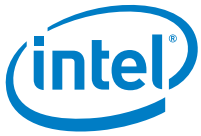
3 Frames May Be Dropped with a High Volume of Mixed Traffic Types

Problem:

Under certain highly restricted traffic conditions a very low rate of frame drops may be observed. For frame drops to occur, at least 14 ports must be active with minimum size, or near minimum size frames, all at maximum frame rate, and the traffic must consist of mixed frame types such that the frame header lengths have a high degree of variability (e.g., layer 2 and IPv4 and IPv6 frames arriving in nearly equal proportion).

Under these conditions, aggregate drop rates may be observed as high as:

- ~0.0008% - with deep inspection enabled
- ~0.000002% - with no deep inspection, but with IPv6
- ~0.0000002% - no L4 parsing, with IPv4 & IPv6
- ~0.0000000003% - IPv4 w/ L4 parsing, no IPv6



- 0% - L2 or IPv4 only, no L4 parsing, no IPv6

Implication:

When a maximum rate of near minimum sized frames occurs, and the traffic consists, for example, of a mixture of IPv4, IPv6 and L2 frames, a very small percentage of frame drops may be seen.

Workaround:

There is no workaround if the traffic conditions match those stated above.

Status: A0-A3=Yes; NoFix

4 Hardware Purge of MAC Table Does Not Work as Expected

Problem:

The hardware purge command per VLAN, per Port or global purges all entries in the MAC Table that match the purge criteria regardless if the MAC entry is static or dynamic. It would be desirable to have the possibility to selectively delete the dynamic entries only or both static and dynamic entries. However, this functionality is not offered.

Implication:

During the time period between deletion of all MAC Table entries and the re-insertion of the static entries (which can take several seconds), a previous (now purged) static MAC Address may be re-learned as a dynamic address. If this occurs, it is never updated to a static address.

Workaround:

The API currently rewrites the static entries automatically after a purge, thus minimizing the time where static entries can create this problem.

This workaround is implemented in API release 2.3.

Status: A0-A3=Yes; NoFix

5 No Way to Disable CM_SHARED_PAUSE_WM per Port

Problem:

If CM_SHARED_PAUSE_WM is used, there is no way to prevent pause frames from being generated on ports that should not pause. The shared PAUSE threshold should only be employed when the switch is used as a lossless fabric where all ports may generate PAUSE.

Implication:

If a non-pausing TX port is above its private WM, and CM_SHARED_PAUSE_WM is triggered from any ingress port, a pause frame is generated on that port.

Workaround:

The work around is to use the CM_RX_SMP_PAUSE_WM pause watermarks instead of the global pause watermarks when PAUSE generation needs to be enabled or disabled per port. PAUSE is generated according to RX port usage instead of a global occupancy level.

Status: A0-A3=Yes; NoFix

6 Canonical GloRT is Applied; GloRT for the Member Port is Expected

Problem:

If a frame is sent in on a LAG destined to the CPU, the ISL tag includes the canonical GloRT instead of the source GloRT.

Implication:

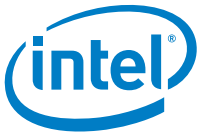
This is incorrect. For a frame sent to the CPU, the source GloRT association should the LAG member GloRT. This is necessary for stack applications such as LACP protocol to operate properly.

Workaround:

The workaround is to associate a USR bit value for each source port in the chip.

On each port, the PORT_CFG_ISL.USR register field is set to be equal to the source GloRT for that port. This allows the software to differentiate between ports on a LAG since the USR field is carried in the ISL tag for the frame.

This workaround is implemented in API release 2.1.



Status: A0-A3=Yes; NoFix

7 Double Tagging Identical VLAN Ethertypes Breaks L3/L4 Parsing

Problem:

To parse L3 information in a frame, the switch must know how to parse the VLAN tags embedded in the frame. When the FM4000 receives a double-tagged frame where the Ethertypes of both VLAN tags are the same (if the SVLAN and CVLAN are configured with the same Ethertype) the parser always considers the second VLAN tag as an Ethertype and the frame is processed as L2.

Implication:

This means that the customer and service Ethertypes cannot be the same, since the service Ethertypes are always interpreted as customer VLAN Ethertypes. If they are the same, the device incorrectly parses the second VLAN Ethertype as the TYPE field of the frame, considers this frame L2 only and stops parsing any further.

Workaround:

In this configuration, L3/L4 information can be read into the FFU using L2 deep inspection. This allows support for L3/L4 ACLs, IGMP snooping and L3 priority association. It does not support L3 frame modification functions such as IP routing. Sample application code for this workaround is available upon request.

Status: A0-A3=Yes; NoFix

8 TCN Precedence is Incorrect for Learning or Aging Events

Problem:

A security, parity or binfull event may mask a learning or aging event, causing the software and hardware tables to lose synchronization.

Implication:

After every TCN error event (security, parity and binfull), the entire MAC Table must be scanned for inconsistencies. If it is not necessary to receive notification of these types of events, the software can set the ErrorEventsLimit to 0b to eliminate this problem.

Workaround:

The software work around is to do a full polling on every security, binfull, or parity event because it is unknown if a MA write TCN entry was masked.

This workaround is implemented in API release 2.1.

Status: A0-A3=Yes; NoFix

9 Strict/Round-Robin Scheduler Bug

Problem:

Configuring egress scheduling with Deficit Round-Robin (DRR) weights of less than 2x max frame size causes unfairness between queues being scheduled using round-robin scheduling.

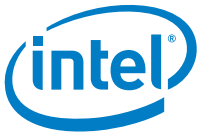
Implication:

This limits the minimum queue size allowable for correct operation. Larger queue weights mean that larger bursts of frames within a single scheduling group are dequeued than otherwise without this limitation.

Workaround:

When using DRR, the minimum queue weight used should be 2x max frame size.

Status: A0-A3=Yes; NoFix



10 RX Rate Limiter Drop Threshold Quanta

Problem:

A value of 1 is interpreted as 1 byte not 256 bytes as stated in the *Intel® Ethernet Switch FM4000 Datasheet* for CM_RX_RATE_LIM_THRESHOLD.drop register field. The datasheet lists the CM_RX_RATE_LIM_THRESHOLD.off register field correctly.

Implication:

This has the effect of shrinking the burst absorption capacity of the token bucket rate limiter by a factor of 256, limiting the maximum burst that the rate limiter tolerates before rate limiting traffic.

Workaround:

A value of 0xFFFF is 65 KB, which corresponds to 1024 pause quanta. This is enough margin to not drop frames.

Status: A0-A3=Yes; NoFix

11 tcnCountErrorEvents Becomes Invalid, Blocking New Error Events

Problem:

When an error event is loaded into the TCN FIFO, the current error count is incremented and stored with it. When the entry is removed, the error count is reduced by the value stored in the FIFO instead of just decrementing the counter.

Implication:

This can cause the counter to decrement below zero to 0x1FF, so no more events are stored in the FIFO, even if the ErrorEventsLimit is set to the maximum of 0x1FF.

Workaround:

The workaround is to periodically issue a MA_PURGE of an unused GloRT. This updates the counter so it does not stay in the bad state forever. The error limit should also be set to the max (0x1FF).

Status: A0-A3=Yes; NoFix

12 CPU May Encounter Very Long Bus Cycles When Sending Frames to the Switch

Problem:

When the local CPU sends a frame, the switch holds the bus cycle if the internal port 0 logic is busy processing another frame from another source (for example in-band management or congestion notifications). If this other frame is a large in-band management frame, the local CPU bus cycle could be stalled enough to eventually exceed the CPU bus timer. This causes the CPU operation to abort, leaving the switch CPU interface in an unknown state.

Implication:

Systems looking to combine in-band management with an attached CPU must take care to limit the duration of in-band management operations such that they do not exceed the maximum time allotted for a CPU bus transaction.

Workaround:

In this case, either the bus timeout should be increased or the maximum length of in-band management frames should be reduced.

Status: A0-A3=Yes; NoFix

13 Reserved Frames Similar to PAUSE Are Not Dropped Correctly

Problem:

According to IEEE Std 802.3-2005 - sub-clauses 3.5, 31.4, 31.5.1, and 31.5.2, the device should drop any frame that has either a 0x8808 Ethertype or a destination MAC Address set to 0x0180c2000001. The device only drops frames that match both the Ethertype and destination MAC Address.

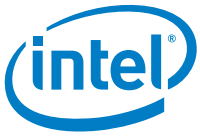
Implication:

Frames with either a matching PAUSE Ethertype or a matching PAUSE DMAC are not dropped unless both the Ethertype and DMAC match.

Workaround:

The workaround is to add a trigger to drop all frames that match the Ethertype, and add either an ACL or entry in the MAC Table (in each FID) to drop the frames with the specific MAC Address.

Status: A0-A3=Yes; NoFix



14 IP Multicast Packets Are Counted as Flood Packets in Group 6

Problem:

If the FFU is used to switch/route IP multicast packets and the destination MAC Address/VLAN is not an entry in the MAC Table, the packets are counted as flooded in group 6.

Implication:

Packets are counted as flooded, while the packets are actually multicast and not flooded.

Workaround:

None.

Status: A0-A3=Yes; NoFix

15 SGMII Auto-Negotiation Reports Failure Even When Successful

Problem:

The device supports SGMII/1000Base-X auto-negotiation as defined in 802.3 clause 37. The SGMII logic re-purposes the NextPage bit to LinkUp status. The device was not designed to support this particular bit at this location, and this causes the device to wait for the next page, which never comes.

Implication:

As a result, the auto-negotiation state machine times-out rather than terminating gracefully. However, the SGMII capability announcement is still received correctly.

Workaround:

The software fix considers the auto-negotiation to have completed successfully even when the AN auto-negotiation state machine times out for SGMII.

This software workaround is implemented in API release 2.5.1.

Status: A0-A3=Yes; NoFix

16 Parity Error Interrupts Are Not Reported in LSM_INT_DETECT

Problem:

There is an error in the implementation of the interrupt pin logic. Bits 1 to 31 of SW_IP and PERR_IP are not reported at the LSM_INT_DETECT pin, but these register bits do report the correct status of the interrupts and are clearable using the usual mechanism.

Implication:

The interrupt register values operate correctly, but not the interrupt pin LSM_INT_DETECT.

Workaround:

The software workaround is to poll PERR_IP to detect interrupt events on bits 1 to 31. These interrupts are expected to occur only in rare error cases so only very low polling rates are needed (~1-10s would be fine).

Status: A0-A3=Yes; NoFix

17 CRC Check Not Done on Truncated Frames in Cut-Through Mode

Problem:

In cut-through mode, when the device makes truncated copies of frames, it truncates them and gives them a correct CRC even if the original CRC was bad.

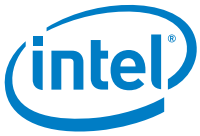
Implication:

These frames may not be flagged for errors downstream. If the purpose of truncation is just for network monitoring, this probably does not matter.

Workaround:

Disable truncation of mirrored frames.

Status: A0-A3=Yes; NoFix



18 EBI Timing Specification Needs to Be Changed

Problem:

The device was modified to move the RXRDY one cycle earlier, which violates the EBI timing specification in the *Intel® Ethernet Switch FM4000 Datasheet* by 0.7 ns. The current maximum timing from clock to output valid is 3.5 ns. The new specification is 4.2 ns.

Implication:

This affects board-level timing.

Workaround:

None.

Status: A0-A3=Yes; NoFix

19 Parse Errors Cannot Be Disabled

Problem:

In each port, the MAC_CFG_2 bit 2 that disables the flagging of a frame as having a parse error is not connected to anything, so it is currently impossible to disable this feature through the EPL.

Implication:

All parse errors are flagged.

Workaround:

The workaround is to install a trigger that un-drops frames with parse errors. The trigger can be configured to un-drop only on certain ports, replicating the functionality of the per-port configuration bit.

Status: A0-A3=Yes; NoFix

20 TCN FIFO Has MA_TABLE Parity Error Bit Set When Soft Learning is On

Problem:

When soft learning is on, the TCN FIFO can return a MA_TABLE entry that has the parity error bit set.

Implication:

This has the potential to confuse software, but is not considered a significant issue since the bit does not give any useful information in this case.

Workaround:

Software should ignore the parity bit.

Status: A0-A3=Yes; NoFix

21 Multi-Chip TX Mirroring Requires Dedicated Inter-Switch Link on Mirroring Switch

Problem:

During TX mirroring, mirrored traffic is copied to a specific output port on the mirroring chip. The TX EPL configuration on this output port must match that of the mirroring port for the copies to have the same frame format (VLAN tag, IP header, etc). If the monitor port for this TX mirror lies on a different chip in a multi-chip system, these mirror copies are GloRT forwarded to the monitor port once they are replicated on the output port.

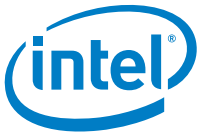
Implication:

In these cases, the output port cannot be a standard inter-switch link, since inter-switch links are pre-configured to have a specific TX EPL configuration that cannot be changed.

Workaround:

In a multi-chip system, if the mirror port and the monitor port are on different chips, an inter-switch link port on the mirroring chip must be dedicated to the TX mirroring function. If only one inter-switch link is available, TX mirroring to a remote monitor port cannot be supported.

Status: A0-A3=Yes; NoFix



22 The Datasheet Description of Register CM_GLOBAL_WM is Insufficient

Problem:

The CM_GLOBAL_WM register needs to be set sufficiently below the top of memory so that the device never requests more memory than it has. Requesting more memory than what is available can cause frame corruption, or potentially chip deadlock.

Implication:

Since memory use is accounted for on the tail of each frame, you need to give a "skidpad" of one maximum frame size (rounded up to the nearest segment) for each port. This guarantees correctness, accounting for the very improbable scenario that you get a max size frame of each port all perfectly aligned right at the moment where the memory use crosses the global WM.

Workaround:

This WM is managed by the API, and is adjusted when the maximum frame size of a port is changed.

This workaround is implemented in API release 2.3.

Status: A0-A3=Yes; NoFix

23 TX Polarity Reversal is Per-Port Not Per-Lane

Problem:

The polarity of the TX SerDes can only be inverted per-port, not per-lane as the spec originally intended. The RX polarity can be inverted per lane.

Implication:

Board designers should not assume that the TX polarity of the SerDes can be inverted per lane.

Workaround:

None.

Status: A0-A3=Yes; NoFix

24 Idle Disparity Problem with Lane Reversal in 1 GbE mode

Problem:

When lane reversal is used in 1 GbE mode on the transmitter (i.e. when the FM4000 uses lane D for transmission in SGMII or 1000BaseX mode) the transmitter fails to balance disparity at the end of the frame. It should do this by optionally transmitting an /I1/ idle codeword, but always transmits an /I2/ codeword instead. This is a non-conformance to IEEE clause 36.2.4.12.

Implication:

This may be seen as receive errors when sending packets to other vendor's silicon. With most vendors, there are no errors in packet reception. However, this errata could cause an interoperability problem with other vendors whom incorrectly processes the disparity error.

Workaround:

Recommend to always use lane A in 1 GbE mode for new designs. When using SFP+ PHYs that support SGMII and lane reversal, lane reversal in the PHY device can be used to ensure lane A is used in 1 GbE mode.

Status: A0-A3=Yes; NoFix

25 MTU Checks Applied to All Unicast IP Frames

Problem:

MTU checks are applied to all unicast IP frames, while in reality they should only be applied to routed frames.

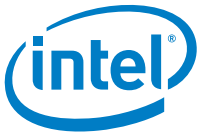
Implication:

All L2 forwarded frames that fit inside max_frame_size but are over the MTU for the ingress interface is trapped to the CPU.

Workaround:

Use a global trigger to un-trap these frames if the ftype == switched. This trigger would need to be placed so that any other trigger based traps/logs would take precedence over this trigger.

Status: A0-A3=Yes; NoFix



26 IP Option Traps Are Applied to All IP Frames

Problem:

IP option traps are applied to all IP frames, while in reality they should only be applied to routed frames.

Implication:

All unicast and multicast IP frames are trapped when using IP option checks, not just routed frames.

Workaround:

Use a global trigger to un-trap these frames if the ftype == switched. This trigger would need to be placed so that any other trigger based traps/logs would take precedence over this trigger.

Status: A0-A3=Yes; NoFix

27 Multiple Triggers with Action 'Drop' Cannot Be Combined

Problem:

When multiple triggers hit a frame, and they each have “drop” actions, then all the DropMasks should be ORed together before being applied to the DestMask:

NewDestMask = ~OR(DropMask[i]) & DestMask

However, the DropMasks are ANDed together instead:

NewDestMask = ~AND(DropMask[i]) & DestMask

If a frame hits a trigger that has the “drop” action enabled (TRIGGER_ACTION_CFG_1.ForwardingAction == 3), learning is disabled for the frame, even if the frame is still forwarded.

In the extreme case, where TRIGGER_ACTION_DROP.DropMask == 0, the frame is not dropped on any ports but learning is still disabled.

Implication:

This makes it difficult to use multiple triggers with the “drop” action. Learning should only be disabled if the drop action causes the frame not to be forwarded out any ports.

Workaround:

This workaround is implemented in API release 2.5.1.

Status: A0-A3=Yes; NoFix

28 In 10/100/1000 Mode, Non Word Aligned Frames May Stretch IFG

Problem:

In 1 GbE mode and its sub-modes 10 MbE and 100 MbE, when transmitting back-to-back frames that are not word aligned (specifically frames whose length modulo 4 is equal to 1 or 2), the TX EPL may add up to two extra idles, increasing the minimum inter-frame gap by 1 or 2 bytes. These extra idles limit data bandwidth of 1 GbE ports if the minimum TX IFG of these ports is set to 12. Sending non-word aligned frames in on one port causes the frames to slowly queue up on egress as the switch is forced to buffer one or two bytes per frame in exchange for the extra bytes of IFG. Sustained traffic incurs greater frame latency and has low amounts of frame loss as the internal buffering becomes full.

Implication:

This could cause high latency due to queue back-pressure or dropped frames.

Workaround:

In 10/100/1000 port speed modes, the TX IFG may be set 10 to allow the switch compensate for the extra idles added in this case. This implies that other frames may be sent out with an IFG of 10 or 11.

Status: A0-A3=Yes; NoFix

29 Certain FIBM Request Message Causes Runt Response Message

Problem:

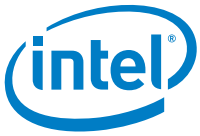
If an FIBM request message requests 7 words, an incorrect response message may be returned with a random value at the end of the message.

Implication:

This could cause incorrect values being returned from a switch register.

Workaround:

The software API version 2.5.1 checks to see if the request results in a response of 7 words. If so, it adds an extra read command to an MSB register to cause the response to be greater than 7 words.



Status: A0-A3=Yes; NoFix

30 Untagged VLAN Tag Frames Do Not Get VID/VPRI Values in ISL Tag

Problem:

When a frame is received on an external port, its VLAN tag information is encapsulated in the VTYPE field of the frame. A VTYPE of zero indicates an untagged frame, while a VTYPE of 1, 2 or 3 represents the type of outer VLAN tag encountered on the frame (CVLAN, SVLANA, and SVLANB respectively). When a frame egresses without a VLAN tag, the VID/VPRI of the egress ISL tag is zeroed out.

Implication:

If this frame is forwarded out an ISL tagged port, the VLAN tagging information is updated given the egress port's VLAN Tag table state. Once the tagging state is updated, the new VTYPE is represented in the ISL tag. If the VTYPE of the frame is zero, the VLAN information (VID and VPRI) in the ISL tag is zero. This is the case if the egress VLAN Tag table sets the VTYPE to zero. This also happens if an ingressing untagged frame is mirrored or trapped out an ISL tagged port such as the CPU.

Workaround:

In a single-chip scenario, this only affects the CPU port, which is always ISL tagged. In this case, to determine the VLAN of incoming untagged frames, the default VLAN of the incoming port is associated with the frame. This workaround has been implemented for all platforms in API release 2.5.2.

In a multi-switch scenario, ISL links should always tag on all active VLANs that include the internal links. The side effect of this requirement is that certain multi-chip RX egress mirrors always produces VLAN tagged mirror copies of frames, if the RX mirror port and RX monitor port are on different chips linked via ISL connections.

Status: A0-A3=Yes; NoFix

31 CPU Port Does Not Pad Certain Frame Sizes During Frame Transmission

Problem:

Frames of length 57, 58, or 59 bytes received from the CPU on the CPU port (port 0) are not padded, and cause rxUndersizedPkts, rxFrameSizeErrors and rxBadOctets to increment. Smaller and larger frames are received on this port without errors.

Implication:

If the CPU port (port 0) receives a frame from the CPU that is 57 to 59 bytes in length, the frame is not padded to 64 bytes, even if the configuration bit MSB_CFG.padHsmFrame is enabled. Frames smaller than 57 bytes are correctly padded, while frames 60 bytes and over do not require any padding.

Workaround:

By default, runt frame padding on the egress port (MAC_CFG_2[egress port].padRuntFrames) ensures that the frame leaves the chip with the minimum frame size. In this case, rxUndersizedPkts and rxFrameSizeErrors counters increment on port 0, which is expected, and can be ignored in this case. If runt frame padding is disabled on an egress port, frames of this size need to be padded to 60 bytes in software.

Status: A0-A3=Yes; NoFix

32 Management Access to FID Tables Not Reliable Under Heavy Traffic Load

Problem:

Reading the INGRESS_FID_TABLE or EGRESS_FID_TABLE may or may not be successful. If a read failure occurs, the resulting data is zero. When writing these tables, the write may or may not be successful. If the write fails, the table is left unaffected. This unreliability only occurs while forwarding traffic, and the probability of failing increases as the traffic rate increases.

Implication:

The software may read incorrect FID values, or may fail to successfully write FID values under heavy traffic load.

Workaround:

The workaround is in three parts:

1. To work around read failures, when a FID table is written, the value written is cached by software. Read operations are satisfied using the cached values.
2. To work around write failures, when a value is written to a FID table, it is then read back to ensure the write was successful (the read-back is from hardware, not the cache from part 1).

   If the write failed, it is repeated until the written value is successfully read back. Since reads are not reliable, the read-back must be repeated until a non-zero value is read.

   Since a successful read of zero data cannot be distinguished from a failed read, a value of zero must never be written to the tables. For the ingress FID table, writing a zero is avoided by always setting bit 1, which is not used by the hardware. Writing a zero to the egress FID table is avoided in part 3.
3. There are no spare bits in the egress FID table with which to avoid writing a zero value as there is for the ingress FID table. To avoid writing a zero to the egress table, whenever a zero value is needed, the FID fields of the corresponding EGRESS_VID_TABLE entries are changed to index a reserved EGRESS_FID_TABLE entry (e.g., 4095) that is never written to, and always contains the hardware default zero value. When the original FID table entry has a non-zero value written to it, the corresponding EGRESS_VID_TABLE entries have their FID fields restored.

This three-part workaround is implemented in API release 2.5.5.

Status: A0-A3=Yes; NoFix

33 An FFU TCAM Read Can Corrupt a Subsequent FFU TCAM Write

Problem:

Reading an FFU TCAM slice can cause a subsequent write with the case bit set to 0b on any of the next few slices to result in the case bit actually being written as a 1b.

Implication:

This may result in incorrect TCAM hits and/or misses on frame traffic.

Workaround:

After reading any TCAM slice and before performing a subsequent TCAM write, slice 31 should be read. Since slice 31 is the last slice, a subsequent write is not corrupted on any slice.

This workaround is implemented in API release 2.5.6.

Status: A0-A3=Yes; NoFix

34 Erroneous Port Linkup in 1 GbE Mode

Problem:

In 1 GbE mode, when a port is brought up with a high traffic rate hitting port from the wire, in some cases the SerDes reports link up but won't pass frames.

Implication:

Port may get stuck and won't pass frames.

Workaround:

Software detects the condition and resets the port until the port starts passing frames.

This workaround is implemented in API release 2.5.6.

Status: A0-A3=Yes; NoFix

35 High Volume of Frame Discards in the Scheduler May Cause the Device to Become Unresponsive

Problem:

If a high volume of frames is scheduled for transmission, and subsequently discarded by the scheduler, it is possible for the device to enter a state where it no longer forwards on some or all enabled ports. Reading from certain frame processing registers (e.g., the MAC Table) while the chip is in this state may cause the management bus to stop responding. In all cases the device requires a full reset to recover and function normally.

This scenario is encountered only when the scheduler is required to discard a high volume of frames that are waiting to be transmitted, due to one or both of the following conditions:

- **Frame Timeout:** Scheduler discards at the required rate can occur if a large number of frames across 12 ports or more are timed out at the same time.
- **TX Error Discard:** Scheduler discards at the required rate can occur if a large number of frames across 12 ports or more with a frame error (such as bad CRC) were not dropped on reception and are held in the memory long enough to be discarded before the start of transmission.

Frames that are discarded using other methods, including drops in the frame processing pipeline and drops by congestion management, cannot create this condition, as these frames are not dropped in the scheduler.

Implication:

If the conditions above are met, the device requires a full reset to function normally.

Workaround:

Hardware-based frame timeouts and error frame discards should be disabled and not used. As of API 2.5.7, these features are disabled by default.

Status: A0-A3=Yes; NoFix

2.5 Fixed Errata (Historical Notes)

1 All /R/ Columns Are Deleted

Status: A0-A2=Yes; NoFix. A3=No; Fixed

This issue is resolve. Related content resides in the *Intel® Ethernet Switch FM4000 Datasheet*.

2 LCI RX_RDY De-Asserted One Cycle Late

Status: A0-A1=Yes; NoFix. A2=No; Fixed

This issue is resolve. Related content resides in the *Intel® Ethernet Switch FM4000 Datasheet*.

3 Statistics Gathering May Lead to Deadlock Under Heavy Load

Status: A0-A1=Yes; NoFix. A2=No; Fixed

This issue is resolve. Related content resides in the *Intel® Ethernet Switch FM4000 Datasheet*.

4 Source Address Lookup May Lead to Deadlock Under Heavy Load

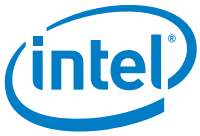
Status: A0-A1=Yes; NoFix. A2=No; Fixed

This issue is resolve. Related content resides in the *Intel® Ethernet Switch FM4000 Datasheet*.

5 FM4000 May Be Irreparably Damaged with Improper Bring-Up Sequence

Status: A0-A1=Yes; NoFix. A2=No; Fixed

This issue is resolve. Related content resides in the *Intel® Ethernet Switch FM4000 Datasheet*.



6 Pause Reception May Operate Unreliably with Routing Enabled

Status: A0-A1=Yes; NoFix. A2=No; Fixed

This issue is resolve. Related content resides in the *Intel® Ethernet Switch FM4000 Datasheet*.

7 Pause Transmission May Operate Unreliably

Status: A0-A1=Yes; NoFix. A2=No; Fixed

This issue is resolve. Related content resides in the *Intel® Ethernet Switch FM4000 Datasheet*.

8 MAC Address Table Change Notifications May Cease at High Temperature

Status: A0-A2=Yes; NoFix. A3=No; Fixed

This issue is resolve. Related content resides in the *Intel® Ethernet Switch FM4000 Datasheet*.

9 Source MAC Address Migration May Not Be Reported Reliably

Status: A0-A1=Yes; NoFix. A2=No; Fixed

This issue is resolve. Related content resides in the *Intel® Ethernet Switch FM4000 Datasheet*.

10 Background MAC Address Table Management Access May Freeze Aging

Status: A0-A1=Yes; NoFix. A2=No; Fixed

This issue is resolve. Related content resides in the *Intel® Ethernet Switch FM4000 Datasheet*.

11 High Level of Transitions in Frame Handler May Cause Unpredictable Behavior

Status: A0-A1=Yes; NoFix. A2=No; Fixed

This issue is resolve. Related content resides in the *Intel® Ethernet Switch FM4000 Datasheet*.

12 Tx/Rx Queue Bottleneck May Lead to Deadlock

Status: A0-A2=Yes; NoFix. A3=No; Fixed

This issue is resolve. Related content resides in the *Intel® Ethernet Switch FM4000 Datasheet*.

13 SPI MOSI/MISO Pins Reversed (897-Ball Package Only)

Status: A0-A2=Yes; NoFix. A3=No; Fixed

This issue is resolve. Related content resides in the *Intel® Ethernet Switch FM4000 Datasheet*.

14 Triggered Forced Learning Not Operational

Status: A0-A1=Yes; NoFix. A2=No; Fixed

This issue is resolve. Related content resides in the *Intel® Ethernet Switch FM4000 Datasheet*.

15 Group 4 and Group 5 Counters Are Unreliable

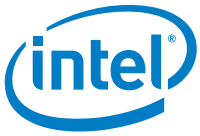
Status: A0-A1=Yes; NoFix. A2=No; Fixed

This issue is resolve. Related content resides in the *Intel® Ethernet Switch FM4000 Datasheet*.

16 PARITY_EVEN Needs Pull-Up or Pull-Down

Status: A0-A1=Yes; NoFix. A2=No; Fixed

This issue is resolve. Related content resides in the *Intel® Ethernet Switch FM4000 Datasheet*.



17 VLAN Counters Do Not Clear on Write

Status: A0-A2=Yes; NoFix. A3=No; Fixed

This issue is resolve. Related content resides in the *Intel® Ethernet Switch FM4000 Datasheet*.

18 Power-Up Supply Sequence Must Be Observed

Status: A0-A2=Yes; NoFix. A3=No; Fixed

This issue is resolve. Related content resides in the *Intel® Ethernet Switch FM4000 Datasheet*.

§ §

NOTE: ***This page intentionally left blank.***